Optimal decentralized control of coupled subsystems with control sharing

Aditya Mahajan

August 9, 2012

Abstract

Subsystems that are coupled due to dynamics and costs arise naturally in various communication applications. In many such applications the control actions are shared between different control stations giving rise to a *control sharing* information structure. Previous studies of control-sharing have concentrated on the linear quadratic Gaussian setup and a solution approach tailored to continuous valued control actions. In this paper a three step solution approach for finite valued control actions is presented. In the first step, a person-by-person approach is used to identify redundant data or a sufficient statistic for local information at each control station. In the second step, the common-information based approach of Nayyar et al. (2011) is used to find a sufficient statistic for the common information shared between all control stations and to obtain a dynamic programming decomposition. In the third step, the specifics of the model are used to simplify the sufficient statistic and the dynamic program. As an example, an exact solution of a two-user multiple access broadcast system is presented.

I. INTRODUCTION

A. Motivation

In this paper, we investigate a modular architecture for networked control systems that consists of a collection of dynamically coupled subsystems, each with a local control station. Each control station observes, either fully or partially, the state of its subsystem, but does not observe the state of other subsystems.[1] In addition, each control station observes the control action of all other

Aditya Mahajan is with the Department of Electrical and Computer Engineering, McGill University, Montreal, QC H3A 2A7, Canada. `aditya.mahajan@mcgill.ca`

[1]In the formal description of the model Section II-A, we also assume that, in addition to the local observations, the control stations also observe a shared state. With this slight generalization, the model can capture more applications, but does not add any additional conceptual difficulties. For that reason, we do not include the shared state in the discussion in this section.

control stations with one-step delay. Such a *control sharing* happens naturally in applications like multi-access broadcast [1], [2] (see Section VI), paging and registration in mobile cellular systems [3], real-time communication with feedback [4], and sensor networks [5].

Each control station affects the state evolution of each substation; thus the substation have coupled dynamics. The per unit cost depends on the state of all subsystems and the control actions of all control stations; thus the control stations are coupled through cost. No control station knows the information available to other control stations. Hence, the system has a non-classical information structure [6], [7].

Each control station has a perfect recall, that is, it chooses a control action based on the history of its observations and control actions. Since the domain of the control laws increases with time, we need to find a time-homogeneous sufficient statistic for the past data at each controller to pose and solve the infinite horizon optimal control problem. Finding such sufficient statistics is difficult due to the non-classical nature of information. For systems with classical information, a sufficient statistic at a control station captures the affect of past data (at that control station) on future estimation (at that control station). This feature is called the *dual affect* of control. For systems with non-classical information, in addition to the above, a sufficient statistic at a control station must capture the affect of past control actions (at that control station) on the future estimation *at other control stations*. This feature is called the *triple affect* of control—the third affect being the *signaling* affect. The control sharing information structure makes the signaling affect explicit; as such solution techniques to control sharing provide insights to other non-classical information structures where the signaling affect is implicit.

B. Literature Overview

There are only a few general frameworks of dynamic programming for systems with non-classical information structure: the sequential team approach for finite horizon systems [8], a common-information based approach for finite horizon systems [9], and a two-step solution approach for two-agent finite and infinite horizon systems [10]. We are interested in a solution framework that works for multiple control stations and extends to infinite horizon systems, and hence, these generic dynamic programming approaches are not applicable.

Most of the research on non-classical information structure has focused on specific system-dynamics and/or specific information structures. We briefly describe some of these approaches

below (see [11] for a detailed discussion.)

The special case of linear dynamics and Gaussian disturbances with specific information structures has received considerable attention in the literature. Examples include static teams [12], [13], partially nested teams [14], [15], stochastically nested teams [16], and quadratic invariant teams [17]–[19] We are interested in systems with non-linear dynamics. More importantly, the control-sharing information structure is neither static, nor partially nested, nor stochastically nested, nor quadratically invariant.

The special case of non-classical information structure with specific data sharing between the control stations has also received considerable attention in the literature. Examples include delayed-state observation [20], delayed (observation) sharing [6], [21], [22], control sharing [23], [24], periodic sharing [25], belief sharing [16], and partial history sharing [26]. Out of these, the models closest to our setup are control-sharing and partial history sharing.

As described earlier, in a control-sharing information structure, control stations can directly signal to one another through their control actions. This signaling aspect was exploiting in [23], [24] by explicitly embedding the local observations in the control actions with arbitrary small perturbation of the control action. Their embedding technique relies on: (i) real-valued random variables have infinite information (in an information theoretic sense); and (ii) the existence of measurable bijections between Euclidean spaces. Such an embedding of observations converts the control sharing information structure to a one-step delayed (observation) sharing information structure, which is also a partially nested information structure. Then, the solution techniques for partially nested teams give an approximate solution for the control-sharing information structure [24]. However, our motivation for investigating these models comes from communication networks, most of which have finite valued control actions.[2] Embedding observations in finite valued control actions is not possible. Hence, the solution technique of [23], [24] does not work for finite valued action spaces.

In a system with control sharing, each control station knows part of the history of data at all control stations. Thus, a control-sharing information structure is also a *partial history sharing* information structure, for which the following solution approach is proposed in [26]. Split the

---

[2]Even otherwise, the assumption of noiseless sharing continuous valued control actions is not realistic in communication applications because it requires infinite capacity communication channels.

data available at each control station into two parts: a common information part that is commonly shared amongst all control stations, and a local information part that consists of the remaining data. Then, the decentralized stochastic control problem is equivalent to a centralized centralized stochastic control problem in which a fictitious *coordinator* observes the common information and chooses functions that map the local information at each control station to its action. This solution approach extends to infinite horizons only when the location information is not increasing with time—which is not the case in the above model because the local information, which is the history of local state observations, is increasing with time. Hence, the solution approach of [26] is not directly applicable to control sharing information structures.

C. Contributions of the paper

One of the main difficulties in obtaining dynamic programming decomposition for decentralized stochastic control is to identify sufficient statistics (for each control station) that summarize the affect of the history of their observations and actions on future observations and cost. In this paper, we present a three step approach to find such sufficient statistics for decentralized control of dynamically coupled subsystems with control sharing.

In the first step, we use a person-by-person approach and identify either irrelevant data or a sufficient statistic for part of the data at each control station. In the second step, we use the common information approach of [26] and identify sufficient statistic for the common information at all control stations. In the third step, we use the salient features of the model—full or partial observation of local states, dynamic coupling using control actions, and sharing of control actions—to simplify the sufficient statistic obtained in the second step. Using the sufficient statistics of the second step (and their simplification in the third step), we obtain a dynamic programming decomposition which can be extended to infinite horizon discounted cost setup. Such a dynamic programming decomposition is not possible by using either the person-by-person approach or the common-information approach alone.

We use the proposed solution approach to obtain a dynamic programming decomposition for a multiuser broadcast channel, and analytically solve the dynamic program when both users have the same arrival rates. Although this example is very well studied, this is the first result that provides a dynamic programming decomposition for this model.

The rest of this paper is organized as follows. We present two models for coupled subsystems

with control sharing in Section II: the full and partial observation models. We present the above described three step approach for the full observation model in Section III and for the partial observation model in Section IV. We present an example of a two-user multiaccess broadcast channel in Section VI and conclude in Section VII.

D. Notation

Random variables are denoted with upper case letters (X, Y, etc.), their realization with lower case letters (x, y, etc.), and their space of realizations by script letters ($\mathcal{X}$, $\mathcal{Y}$, etc.). Subscripts denote time and superscripts denote the subsystem; e.g. X_t^i denotes the state of subsystem i at time t. The short hand notation $X_{1:t}^i$ denotes the vector $(X_1^i, X_2^i, \ldots, X_t^i)$. Bold face letters denotes the collection of variables at all subsystems; e.g., $\mathbf{X}_t$ denotes $(X_t^1, X_t^2, \ldots X_t^n)$. The notation $\mathbf{X}_t^{-i}$ denotes the vector $(X_t^1, \ldots, X_t^{i-1}, X_t^{i+1}, \ldots, X_t^n)$.

$\Delta(\mathcal{X})$ denotes the probability simplex on the space $\mathcal{X}$. $\mathbb{P}(A)$ denotes the probability of an event A, and $\mathbb{E}[X]$ denotes the expectation of a random variable X. $\mathbb{1}[x = y]$ denotes the indicator function of the statement $x = y$, i.e., $\mathbb{1}[x = y] = 1$ if $x = y$ and 0 otherwise. Let $\mathbb{N}$ denote the set of natural numbers and $\mathbb{Z}_+$ denote the set of non-negative integers.

II. COUPLED SUBSYSTEMS WITH CONTROL SHARING

A. Model and Problem Formulation

System components: Consider a discrete-time networked control system with n subsystems. The state (Z_t, X_t^i) of subsystem i, $i = 1, \ldots, n$, has two components: a *local state* $X_t^i \in \mathcal{X}^i$ and a *shared state* $Z_t \in \mathcal{Z}$, which is identical for all subsystems. The initial shared state Z_1 has a distribution P_Z. Conditioned on the initial shared state Z_1, the initial local state of all subsystems are independent; initial local state X_1^i is distributed according to $P_{X^i|Z}$, $i = 1, \ldots, n$. Let $\mathbf{X}_t := (X_t^1, \ldots, X_t^n)$ denote the local state of all subsystems.

A control station is co-located with each subsystem. Let $U_t^i \in \mathcal{U}^i$ denote the control action of control station i and $\mathbf{U}_t := (U_t^1, U_t^2, \ldots, U_t^n)$ denote the collection of all control actions.

System dynamics: The shared and the local state of each subsystems are coupled through the control actions; the shared state evolves according to

$$Z_{t+1} = f_t^0(Z_t, \mathbf{U}_t, W_t^0) \tag{1}$$

while the local state of subsystem i, $i = 1, \ldots, n$, evolves according to:

$$X^i_{t+1} = f^i_t(Z_t, X^i_t, \mathbf{U}_t, W^i_t) \tag{2}$$

where $W^i_t \in \mathcal{W}^i$, $i = 0, 1, \ldots, n$, is the plant disturbance with distribution P_{W^i}. The processes $\{W^i_t, t = 1, \ldots\}$, $i = 0, 1, \ldots, n$, are assumed to be independent across time, independent of each other, and also independent of the initial state $(Z_1, \mathbf{X}_1)$ of the system.

Note that the updated local state of subsystem i depends only on the previous local state of subsystem i and previous shared state but is controlled by all control stations.

Observation models and information structures: We consider two observation models that differ in the observation of the location state X^i_t at control station i. In the first model, called *full observation model*, control station i perfectly observes the local state X^i_t; in the second model, called *partial observation model*, control station i observes a noisy version $Y^i_t \in \mathcal{Y}^i$ of the local state X^i_t given by

$$Y^i_t = \ell^i_t(X^i_t, \tilde{W}^i_t) \tag{3}$$

where $\tilde{W}^i_t \in \tilde{\mathcal{W}}^i_t$ is the observation noise with distribution $P_{\tilde{W}^i}$. The processes $\{\tilde{W}^i_t, t = 1, \ldots, \}$, $i = 1, \ldots, n$ are assumed to be independent across time, independent of each other, independent of $\{W^i_t, t = 1, \ldots, n\}$, and independent of the initial states $(\mathbf{X}_1, Z_1)$.

In both models, in addition to the local measurements of the state of its subsystem, each control station perfectly observes the shared state Z_t and the one-step delayed control actions $\mathbf{U}_{t-1}$ of all control stations. The control stations perfectly recall all the data they observe. Thus, in the full observation model, control station i chooses a control action according to

$$U^i_t = g^i_t(Z_{1:t}, X^i_{1:t}, \mathbf{U}_{1:t-1}) \tag{4}$$

while in the partial observation model, it chooses a control action according to

$$U^i_t = g^i_t(Z_{1:t}, Y^i_{1:t}, \mathbf{U}_{1:t-1}). \tag{5}$$

The function g^i_t is called the *control law* of control station i. The collection $\mathbf{g}^i := (g^i_1, g^i_2, \ldots, g^i_T)$ of control laws at control station i is called the *control strategy of control station i*. The collection $\mathbf{g} := (\mathbf{g}^1, \mathbf{g}^2, \ldots, \mathbf{g}^n)$ of control strategies of all control stations is called the *control strategy of the system*.

Cost and performance: At time t, the system incurs a cost $c_t(Z_t, \mathbf{X}_t, \mathbf{U}_t)$ that depends on the shared state, the local state of all subsystems, and the actions of all control stations. Thus, the subsystems are also coupled through cost.

The system runs for a time horizon T. The performance of a control strategy $\mathbf{g}$ is measured by the expected total cost incurred by that strategy, which is given by

$$J(\mathbf{g}) := \mathbb{E}\Big[\sum_{t=1}^{T} c_t(Z_t, \mathbf{X}_t, \mathbf{U}_t)\Big] \tag{6}$$

where the expectation is with respect to a joint measure of $(Z_{1:T}, \mathbf{X}_{1:T}, \mathbf{U}_{1:T})$ induced by the choice of the control strategy $\mathbf{g}$.

We are interested in the following optimal control problem:

Problem 1: Given the distributions P_Z, $P_{X^i|Z}$, P_{W^i}, $P_{\tilde{W}^i}$ of the initial shared state, initial local state, plant disturbance of subsystem i, and observation noise of subsystem i (for the partial observation model), $i = 1, \ldots, n$, a horizon T, and the cost functions c_t, $t = 1, \ldots, T$, find a control strategy $\mathbf{g}$ that minimizes the expected total cost given by (6).

B. Applications in communication networks

Control-sharing information structure arises naturally in communication networks, as is illustrated by some applications described below.

1) Paging and registration in cellular networks: Consider a mobile cellular network consisting of two controllers: a network operator and a mobile station. The local state X_t^1 of the network operator is a constant and the local state X_t^2 of the mobile station is its current location that changes in a Markovian manner. We will describe the shared state Z_t later. The control action U_t^1 of the network operator is a permutation of $\mathcal{X}^2$, the set of all possible locations of the mobile, and denotes the order in which mobile station will be searched if there is a paging request. The control action U_t^2 of the mobile station is either X_t^2 (indicating that the mobile station registers with the network) or NR (indicating the mobile station does not register).

At each time, the network may get an exogenous *paging* request to seek the location of the mobile station. If a paging request is received (denoted by $P_t = 1$), the cost of searching is given by the index $i(X_t^2, U_t^1)$ of X_t^2 in U_t^1. If no paging request is received (denoted by $P_t = 0$) and the mobile station registers with the network, a registration cost of r is incurred. The process P_t is a binary-valued Markov process. If either the mobile station is paged or the mobile station

registers with the network, the network operator learns the current location of the mobile station. Let M_t denote the time since the last paging or registration request and S_t denote the location of the mobile station at that time. Then $Z_t = (P_t, M_{t-1}, S_{t-1})$ is the shared state of the system.

The above model corresponds to the model of paging and registration in mobile cellular network considered in [3]. The control action U_t^1 of the network is based on information known to the mobile station, hence U_t^1 is effectively observed at the mobile station. The control action U_t^2 of the mobile station is communicated to the network operator. Hence, this system has control-sharing information structure.

2) Real-time communication: Consider a real-time communication system consisting of an encoder and a decoder. The encoder observes a first-order Markov source S_t. The local state X_t^1 of the encoder is (S_{t-1}, S_t) and the local state X_t^2 of the decoder is a constant. The shared state Z_t is also a constant. The control action U_t^1 of the encoder is a quantization symbol that is communicated to the decoder. The control action U_t^2 of the decoder is an estimate of the one-step delayed source S_{t-1} of the encoder. The cost at each time is given by a distortion between S_{t-1} and U_t^2.

The above model corresponds to the model of real-time communication considered in [27] (specialized to infinite memory). The control action U_t^1 of the encoder is communicated to the decoder. The control action U_t^2 of the decoder is based on the information known to the encoder, hence U_t^2 is effectively observed at the encoder. Hence, this system has the full observation model considered above with shared state $Z_t = \emptyset$.

3) Multiaccess broadcast: Consider a two-user multiaccess broadcast system. At time t, $W_t^i \in \{0, 1\}$ packets arrive at each user according to independent Bernoulli processes with $\mathbb{P}(W_t^i = 1) = p^i$, $i = 1, 2$. Each user may store only $X_t^i \in \{0, 1\}$ packets in a buffer. If a packet arrives when the user-buffer is full, the packet is dropped.

Both users may transmit $U_t^i \in \{0, 1\}$ packets over a shared broadcast medium. A user can transmit only if it has a packet, thus $U_t^i \leq X_t^i$. If only one user transmits at a time, the transmission is successful and the transmitted packet is removed from the queue. If both users transmit simultaneously, packets "collide" and remain in the queue. Thus, the state update for user 1 is given by $X_{t+1}^1 = \max(X_t^1 + U_t^1 \cdot (1 - U_t^2) + W_t^1, 1)$. The state update rule for user 2 is symmetric dual of the above.

Instead of costs, it is more natural to work with rewards in this example. The objective is

to maximize throughput, or the number of successful packet transmissions. Thus, the per unit reward is $c(\mathbf{x}, \mathbf{u}) = u^1 \oplus u^2$, where $\oplus$ means binary XOR.

When the arrival rates at both users are the same ($p^1 = p^2$), the above model corresponds to the two-user multiaccess broadcast system considered in [1], [2], [28]. Slight variation of the above model were considered in [29], [30]. In recent years, the two-user multiaccess broadcast system with asymmetric arrivals ($p^1 \neq p^2$) has been used as a benchmark problem for decentralized stochastic control problems in the artificial intelligence community [31]–[35].

Due to the broadcast nature of the communication channel, each user observes the transmission decision of the other user. Hence the system has the full observation model considered above with shared state $Z_t = \emptyset$. We will revisit this model in Section VI.

III. MAIN RESULT FOR THE FULL OBSERVATION MODEL

In this section, we derive structure of optimal control laws and a dynamic programming decomposition for the full observation model. As stated in the introduction, the full observation model has a partial history sharing information structure [26]. Nayyar *et al.* proposed a *common information based* approach to design systems with partial information sharing. According to their approach, the design of optimal control strategies is investigated from the point of view of a coordinator that observes the shared common information. In the full observation model, the shared common information $C_t = (Z_{1:t}, \mathbf{U}_{1:t-1})$, and the private local information is $L_t^i = \{X_{1:t}^i\}$. According to [26], the posterior probability $\mathbb{P}(Z_t, \mathbf{L}_t \mid C_t)$ is a sufficient statistic for the shared common information C_t.

However, directly using the above approach is not useful for the full observation model because the local information L_t^i at control station i, $i = 1, \ldots, n$, is increasing with time, which causes the dimension of the sufficient statistic $\mathbb{P}(Z_t, \mathbf{L}_t \mid C_t)$ to increases with time; and therefore, $\mathbb{P}(Z_t, \mathbf{L}_t \mid C_t)$ does not work as a sufficient statistic for infinite horizon setup.

In this paper, we present the following three step approach to simplify the structure of the control laws and derive a dynamic programming decomposition (that extends to the infinite horizon setup).

1) Use a person-by-person approach to show that the past values of the local state $X_{1:t-1}^i$ are irrelevant at control station i at time t. Thus, for any control strategy of control station i that

uses $(X^i_{1:t}, Z_{1:t}, \mathbf{U}_{1:t-1})$, we can choose a control strategy that uses only $(X^i_t, Z_{1:t}, \mathbf{U}_{1:t-1})$ without any loss in performance.

2) When attention is restricted to control strategies of the form derived in Step 1, the local information $L^i_t = \{X^i_t\}$ at control station i does not increase with time. Thus, using the results of [26], we can show that $\Pi_t = \mathbb{P}(\mathbf{X}_t, Z_t \mid C_t)$ is a sufficient statistic for the common information C_t and is also an information state for dynamic programming.

3) Using the system dynamics, show that Π_t defined in Step 2 is equivalent to $(Z_t, \mathbf{\Theta}_t)$, where $\mathbf{\Theta}_t = (\Theta^1_t, \ldots, \Theta^n_t)$ and $\Theta^i_t = \mathbb{P}(X^i_t \mid C_t)$. Using this equivalence, we can simplify the structural result and dynamic programming decomposition of Step 2.

Now, we describe each of these steps in detail. For simplicity of exposition, we assume that $\mathcal{Z}$, $\mathcal{X}^i$, $\mathcal{U}^i$, and $\mathcal{W}^i$, $i = 1, \ldots, n$, are finite. The results extend to general alphabets under suitable technical conditions (similar to those for centralized stochastic control [36]).

Step 1: Shedding of irrelevant information

In this section, we show that the past values of local state $X^i_{1:t-1}$ are irrelevant at control station i at time t, $i = 1, \ldots, n$. In particular:

Proposition 1: In the full observation model, restricting attention to control laws of the form

$$U^i_t = \bar{g}^i_t(X^i_t, Z_{1:t}, \mathbf{U}_{1:t-1}) \tag{7}$$

at all control stations i, $i = 1, \ldots, n$, is without loss of optimality.

A priori, it is not obvious that the past data is irrelevant. Suppose we pick any control station i, $i = 1, \ldots, n$; arbitrarily fix the control strategy of all control stations except station i and consider the subproblem of finding the optimal control strategy at control station i. In principle, the history $X^i_{1:t-1}$ of local states at control station i may give some information about the history $X^j_{1:t-1}$ of local states at control station j, $j \neq i$; and hence, may help in predicting the future control actions of control station j. The following proposition shows that this is not the case. Conditioned on the shared observations $(Z_{1:t}, \mathbf{U}_{1:t})$, the local state processes $\{X^i_t, t = 1, \ldots\}$, $i = 1, \ldots, n$, evolve independently.

Proposition 2: In the full observation model, the local states of all subsystems are conditionally independent given the history of shared state and control actions. Specifically, for any

realization $z_t \in \mathcal{Z}$, $x_t^i \in \mathcal{X}^i$ and $u_t^i \in \mathcal{U}^i$ of X_t^i and U_t^i, $i = 1, \ldots, n$, $t = 1, \ldots, T$, we have

$$\mathbb{P}(\mathbf{X}_{1:t} = \mathbf{x}_{1:t} \mid Z_{1:t} = z_{1:t}, \mathbf{U}_{1:t} = \mathbf{u}_{1:t}) = \prod_{i=1}^{n} \mathbb{P}(X_{1:t}^i = x_{1:t}^i \mid Z_{1:t} = z_{1:t}, \mathbf{U}_{1:t} = \mathbf{u}_{1:t}) \quad (8)$$

See Appendix A for proof. One immediate consequence of the above Proposition is the following:

Lemma 3: Consider the full observation model for an arbitrary but fixed choice of control strategy $\mathbf{g}$. Define $R_t^i = (X_t^i, Z_{1:t}, \mathbf{U}_{1:t-1})$. Then,

1) The process $\{R_t^i, t = 1, \ldots, T\}$ is a controlled Markov process with control action U_t^i, i.e., for any $x_t^i, \tilde{x}_t^i \in \mathcal{X}^i$, $z_t, \tilde{z}_t \in \mathcal{Z}$, $u_t^i, \tilde{u}_t^i \in \mathcal{U}^i$, $r_t^i = (x_t^i, z_{1:t}, \mathbf{u}_{1:t-1})$, $\tilde{r}_t^i = (\tilde{x}_t^i, \tilde{z}_{1:t}, \tilde{\mathbf{u}}_{1:t-1})$, $i = 1, \ldots, n$, and $t = 1, \ldots, T$,

$$\mathbb{P}(R_{t+1}^i = \tilde{r}_{t+1}^i \mid R_{1:t}^i = r_{1:t}^i, U_{1:t}^i = u_{1:t}^i) = \mathbb{P}(R_{t+1}^i = \tilde{r}_{t+1}^i \mid R_t^i = r_t^i, U_t^i = u_t^i)$$

2) The instantaneous conditional cost simplifies as follows:

$$\mathbb{E}[c_t(Z_t, \mathbf{X}_t, \mathbf{U}_t) \mid R_{1:t}^i = r_{1:t}^i, U_{1:t}^i = u_{1:t}^i] = \mathbb{E}[c_t(Z_t, \mathbf{X}_t, \mathbf{U}_t) \mid R_t^i = r_t^i, U_t^i = u_t^i]$$

See Appendix B for proof.

In light of Lemma 3, lets reconsider the subproblem of finding the optimal control strategy for control station i when the control strategy $\mathbf{g}^{-i}$ of all other control stations is fixed arbitrarily. In this subproblem, control station i has access to $R_{1:t}^i$, chooses U_t^i, and incurs an expected instantaneous cost $\mathbb{E}[c_t(\mathbf{X}_t, \mathbf{U}_t) \mid R_{1:t}^i, U_{1:t}^i]$. Lemma 3 implies that the optimal choice of control strategy $\mathbf{g}^i$ is a Markov decision process. Thus, using Markov decision theory [37], we get the following (recall that $R_t^i = (X_t^i, Z_{1:t}, \mathbf{U}_{1:t-1})$ and the choice of $\mathbf{g}^{-i}$ is arbitrary):

Lemma 4: Consider the full observation model for any arbitrary but fixed choice of control strategy $\mathbf{g}^{-i}$ of all control stations except i. Then, restricting attention to control laws of the form

$$U_t^i = g_t^i(X_t^i, Z_{1:t}, \mathbf{U}_{1:t-1}) \quad (9)$$

at control station i is without loss of optimality.

Proof of Proposition 1: Lemma 4 implies that for an arbitrary choice of $\mathbf{g}^{-i}$, control strategies of the form (9) at control station i dominate those of the form (4). Cyclically using the same argument for all control stations proves the result. ∎

Even after shedding $X_{1:t-1}^i$, the data at each control station is still increasing with time. In the next step, we show how to "compress" this data into a sufficient statistic.

Step 2: Sufficient statistic for common data

Consider Problem 1 for the full observation model and restrict control strategies of the form (7). Proposition 1 shows that this restriction is without loss of optimality. We use the results of [26] for this restricted setup.

Split the data at each control station into two parts: the common data $C_t = (Z_{1:t}, \mathbf{U}_{1:t-1})$ that is observed by all control stations and the local (or private) data $L_t^i = X_t^i$ that is observed by only control station i. Note that the common information $C_t \subset C_{t+1}$ is increasing with time, while the local information L_t^i has a fixed size. Thus, the system has *partial history sharing* information structure with finite local memory. Nayyar *et al.* [26] derived structural properties of optimal controllers and a dynamic programming decomposition for such an information structure.

To present the result, we first define the following:

Definition 1: Given any control strategy $\bar{\mathbf{g}}$ of the form (7), let Π_t, $t = 1, \ldots, T$, denote the posterior probability of $(Z_t, \mathbf{X}_t)$ given the common information C_t; i.e., for any $z \in \mathcal{Z}$ and $x^i \in \mathcal{X}^i$, the component $(z, \mathbf{x})$ of Π_t is given by

$$\Pi_t(z, \mathbf{x}) := \mathbb{P}^{\bar{\mathbf{g}}}(Z_t = z, \mathbf{X}_t = \mathbf{x} \mid C_t).$$

The update of Π_t follows the standard non-linear filtering equation. It is shown in [26] that Π_t is a sufficient statistic for C_t; in particular, we have the following structural result.

Proposition 5 ([26, Theorem 2] applied to model of Proposition 1): In the full observation model, restricting attention to control laws of the form

$$U_t^i = \hat{g}_t^i(X_t^i, \Pi_t) \tag{10}$$

at all control stations i, $i = 1, \ldots, n$, is without loss of optimality.

To obtain a dynamic programming decomposition to find optimal control strategies of the form (10), the following *partially evaluated control laws* were defined in [26]: For any control strategy of the form (10), and any realization π_t of Π_t, let

$$\hat{d}_t^i(\cdot) = \hat{g}_t^i(\cdot, \pi_t)$$

denote a mapping from $\mathcal{X}_t^i$ to $\mathcal{U}_t^i$. When Π_t is a random variable, the above mapping is a random mapping denoted by $\hat{D}_t^i$. Let $\hat{\mathbf{d}}_t = (\hat{d}_t^1, \ldots, \hat{d}_t^n)$ and $\hat{\mathbf{D}}_t = (\hat{D}_t^1, \ldots, \hat{D}_t^n)$. Then optimal control strategies of the form (10) are obtained as follows.

Proposition 6 ([26, Theorem 3] applied to the model of Proposition 1): For any $\pi_t \in \Delta(\mathcal{Z} \times \mathcal{X}^1 \times \cdots \mathcal{X}^n)$, define

$$V_T(\pi_T) = \min_{\hat{\mathbf{d}}_T} \mathbb{E}[c_T(Z_T, \mathbf{X}_T, \mathbf{U}_T) \mid \Pi_T = \pi_T, \hat{\mathbf{D}}_T = \hat{\mathbf{d}}_T] \tag{11}$$

and for $t = T - 1, T - 2, \ldots, 1$,

$$V_t(\pi_t) = \min_{\hat{\mathbf{d}}_t} \mathbb{E}[c_t(Z_t, \mathbf{X}_t, \mathbf{U}_t) + V_{t+1}(\Pi_{t+1}) \mid \Pi_t = \pi_t, \hat{\mathbf{D}}_t = \hat{\mathbf{d}}_t] \tag{12}$$

Let $\hat{\Psi}_t(\pi_t)$ denote the $\arg\min$ of the right hand side of $V_t(\pi_t)$, and $\hat{\Psi}_t^i$ denote the i-th component of $\hat{\Psi}_t$. Then, a control strategy

$$\hat{g}_t^i(x_t^i, \pi_t) \in \hat{\Psi}_t^i(\pi_t)(x_t^i)$$

is optimal for Problem 1 with the full observation model.

Step 3: Simplification of the sufficient statistic

In this step, we use Proposition 2 to simplify the sufficient statistic Π_t used in Step 2, and thereby simplify Propositions 5 and 6. For that matter, we define the following.

Definition 2: Given any control strategy $\hat{\mathbf{g}}$ of the form (10), let Θ_t^i, $t = 1, \ldots, T$, denote the posterior probability of X_t^i given the common information C_t, i.e., for any $x^i \in \mathcal{X}^i$, the component x^i of Θ_t^i is given by

$$\Theta_t^i(x^i) := \mathbb{P}^{\hat{\mathbf{g}}}(X_t^i = x^i \mid C_t).$$

The update of Θ_t^i follows the standard non-linear filtering equation. For completeness, we describe this update below.

Lemma 7: There exists a deterministic function F_t such that

$$\mathbf{\Theta}_{t+1} = F_t(\mathbf{\Theta}_t, Z_{t+1}, \mathbf{U}_t, \hat{\mathbf{D}}_t) \tag{13}$$

The proof follows from the law of total probability and Bayes rule. See Appendix C.

We can now simplify the sufficient statistic Θ_t as follows:

Lemma 8: For any $z \in \mathcal{Z}$, $x^i \in \mathcal{X}^i$, $i = 1, \ldots, n$, the values $(z, \mathbf{\Theta}_t(\mathbf{x}))$ are sufficient to compute $\Pi_t(z, \mathbf{x})$.

Proof: The proof follows directly from the definition of Π_t, Θ_t^i and Proposition 2. Let $C_t = (Z_{1:t}, \mathbf{U}_{1:t-1}$ and consider the component $(z, \mathbf{x})$ of Π_t:

$$\Pi_t(z, \mathbf{x}) \overset{(a)}{=} \mathbb{1}[Z_t = z] \cdot \mathbb{P}(\mathbf{X}_t = \mathbf{x} \mid Z_{1:t}, \mathbf{U}_{1:t-1}) \overset{(b)}{=} \mathbb{1}[Z_t = z] \cdot \prod_{i=1}^n \Theta_t^i(x^i)$$

where (a) follows form the law of total probability and (b) follows from Proposition 2. ∎

By substituting $(Z_t, \boldsymbol{\Theta}_t)$ instead of Π_t in Propositions 5 and 6, we get the following:

Theorem 1 (Structure of optimal controllers): In the full observation model, restricting attention to control laws of the form

$$U_t^i = \tilde{g}_t^i(X_t^i, Z_t, \boldsymbol{\Theta}_t) \tag{14}$$

at all control stations i, $i = 1, \ldots, n$, is without loss of optimality.

For any control strategy of the form (14), and any realization $\boldsymbol{\theta}_t$ of $\boldsymbol{\Theta}_t$, let

$$\tilde{d}_t^i(\cdot) = \tilde{g}_t^i(\cdot, z_t, \boldsymbol{\theta}_t)$$

denote a mapping from $\mathcal{X}_t^i$ to $\mathcal{U}_t^i$. When $\boldsymbol{\Theta}_t$ is a random variable, the above mapping is a random mapping denoted by $\tilde{D}_t^i$. Let $\tilde{\mathbf{d}}_t = (\tilde{d}_t^1, \ldots, \tilde{d}_t^n)$ and $\tilde{\mathbf{D}}_t = (\tilde{D}_t^1, \ldots, \tilde{D}_t^n)$. Then optimal control strategies of the form (10) are obtained as follows.

Theorem 2 (Dynamic programming decomposition): For any $z_t \in \mathcal{Z}$ and $\theta_t^i \in \Delta(\mathcal{X}^i)$, $i = 1, \ldots, n$, define

$$V_T(z_T, \boldsymbol{\theta}_T) = \min_{\tilde{\mathbf{d}}_T} \mathbb{E}[c_T(Z_T, \mathbf{X}_T, \mathbf{U}_T) \mid Z_T = z_T, \boldsymbol{\Theta}_T = \boldsymbol{\theta}_T, \tilde{\mathbf{D}}_T = \tilde{\mathbf{d}}_T] \tag{15}$$

and for $t = T - 1, T - 2, \ldots, 1$,

$$V_t(z_t, \boldsymbol{\theta}_t) = \min_{\tilde{\mathbf{d}}_t} \mathbb{E}[c_t(Z_t, \mathbf{X}_t, \mathbf{U}_t) + V_{t+1}(\Pi_{t+1}) \mid Z_t = z_t, \boldsymbol{\Theta}_t = \boldsymbol{\theta}_t, \tilde{\mathbf{D}}_t = \tilde{\mathbf{d}}_t] \tag{16}$$

Let $\tilde{\Psi}_t(z_t, \boldsymbol{\theta}_t)$ denote the $\arg\min$ of the right hand side of $V_t(z_t, \boldsymbol{\theta}_t)$, and $\tilde{\Psi}_t^i$ denote the i-th component of $\tilde{\Psi}_t$. Then, a control strategy

$$\tilde{g}_t^i(x_t^i, z_t, \boldsymbol{\theta}_t) \in \tilde{\Psi}_t^i(z_t, \boldsymbol{\theta}_t)(x_t^i)$$

is optimal for Problem 1 with the full observation model.

IV. MAIN RESULT FOR THE PARTIAL OBSERVATION MODEL

In this section, we derive structure of optimal control laws and a dynamic programming decomposition for the partial observation model. As in the full observation model, we cannot directly use the results of [26] because the local observations $Y_{1:t}^i$ at each control station are increasing with time. To circumvent this difficulty we follow a three step approach, similar to the one taken for the full observation model, and proceed as follows:

1) Use a person-by-person approach to show that $\Xi_t^i(x) := \mathbb{P}(X_t^i = x \mid Y_{1:t}^i, Z_{1:t}, \mathbf{U}_{1:t-1})$ is a sufficient statistic for the history of local observations at control station i at time t. Thus, for any control strategy of control station i that uses $(Y_{1:t}^i, Z_{1:t}, \mathbf{U}_{1:t-1})$, we can choose a strategy that uses only $(\Xi_t^i, Z_{1:t}, \mathbf{U}_{1:t-1})$ without loss of optimality.

2) Steps 2 and 3 are similar to those of the full observation model with X_t^i replaced by Ξ_t^i.

Now, we describe each of these steps in detail.

Step 1: Sufficient statistic for local observations

In this step, we find a sufficient statistic for the local observations $Y_{1:t}^i$ at control station i. For that matter, we define the following:

Definition 3: Given any control strategy $\mathbf{g}$ of the form (5), let Ξ_t^i, $i = 1, \ldots, n$, $t = 1, \ldots, T$ denote the posterior probability of the local state X_t^i of substation i given all the information $(Y_{1:t}^i, Z_{1:t}, \mathbf{U}_{1:t-1})$ at control station i, i.e., for any $x^i \in \mathcal{X}^i$, the component x^i of Ξ_t^i is given by

$$\Xi_t^i(x^i) := \mathbb{P}^{\mathbf{g}}(X_t^i = x^i \mid Y_{1:t}^i, Z_{1:t}, \mathbf{U}_{1:t-1}) \stackrel{(a)}{=} \mathbb{P}^{\mathbf{g}}(X_t^i = x^i \mid Y_{1:t}^i, Z_{1:t-1}, \mathbf{U}_{1:t-1})$$

where (a) follows from the independence of $\{W_t^0, t = 1, \ldots, T\}$ from $\{W_t^i, t = 1, \ldots, T\}$.

The update of Ξ_t^i follows a non-linear filtering equation as shown below.

Lemma 9: For every i, $i = 1, \ldots, T$, there exist a deterministic functions $\tilde{F}_t^i$ such that

$$\Xi_{t+1}^i = \tilde{F}_t^i(\Xi_t^i, Y_{t+1}^i, Z_t, \mathbf{U}_t). \tag{17}$$

The proof follows from the law of total probability and Bayes rule and is similar to the proof of Appendix C.

The main result of this section is the following:

Proposition 10: In the partial observation model, restricting attention to control laws of the form

$$U_t^i = \bar{g}_t^i(\Xi_t^i, Z_{1:t}, \mathbf{U}_{1:t-1}) \tag{18}$$

at all control stations i, $i = 1, \ldots, n$, is without loss of optimality.

The intuition behind this result is as follows. Arbitrarily fix the control strategies $\mathbf{g}_{-i}$ for all control stations other than i. In the full observation model, $(X_t^i, Z_{1:t}, \mathbf{U}_{1:t-1})$ is a state sufficient for performance evaluation at control station i (Lemma 4). In the partial observation model, component X_t^i of this state is not observed. So, the posterior distribution Ξ_t^i on X_t^i given all the data available at control station i should be a sufficient statistic for X_t^i [38].

To show that the above intuition is true, we need to establish two conditional independence properties.

Proposition 11: Proposition 2 is also true for the partial observation model for an arbitrary but fixed choice of control strategy $\mathbf{g}$ of the form (5).

Proposition 12: In the partial observation model, the posterior probability Ξ_t^i of the local states of all subsystems are conditionally independent given the history of shared state and control actions. Specifically, for any Borel subsets E_t^i of $\Delta(\mathcal{X}^i)$, $\mathbf{E}_t = (E_t^1, \ldots, E_t^n)$, $u_t^i \in \mathcal{U}^i$, $z_t \in \mathcal{Z}$, $i = 1, \ldots, n$ and $t = 1, \ldots, T$, we have

$$\mathbb{P}(\mathbf{\Xi}_{1:t} \in \mathbf{E}_{1:t} \,|\, Z_{1:t} = z_{1:t}, \mathbf{U}_{1:t} = \mathbf{u}_{1:t}) = \prod_{i=1}^{n} \mathbb{P}(\Xi_{1:t}^i \in E_{1:t}^i \,|\, Z_{1:t} = z_{1:t}, \mathbf{U}_{1:t} = \mathbf{u}_{1:t}) \quad (19)$$

These results are proved in Appendices D and E.

An immediate consequence of Proposition 11 and Lemma 9 is the following (see Appendix F for proof).

Lemma 13: Lemma 3 is also true for the partial observation model with R_t^i defined as $(\Xi_t^i, Z_{1:t}, \mathbf{U}_{1:t-1})$.

Proof of Proposition 10: The result of Proposition 10 follows from cyclically repeating an argument similar to the argument after Lemma 3. ∎

Steps 2 and 3: Sufficient statistic for common data and its simplification

Compare Proposition 1 of the full observation model with Proposition 10 of the partial observation model. The posterior probability Ξ_t^i in the latter model plays the role of local state X_t^i in the former model. This suggests that we may follow Steps 2 and 3 of the full observation model in the partial observation model by replacing X_t^i by Ξ_t^i. Following this suggestion, define:

Definition 4: Let $\mathring{\Pi}_t$ denote the posterior probability on $(Z_t, \boldsymbol{\Xi}_t)$ given the common information C_t, i.e., for any $z \in \mathcal{Z}$ and any Borel subsets E^i of $\Delta(\mathcal{X}^i)$ and $\mathbf{E} = (E^1, \ldots, E^n)$,

$$\mathring{\Pi}_t(z, \mathbf{E}) = \mathbb{P}(Z_t = z, \boldsymbol{\Xi}_t \in \mathbf{E} \mid C_t) \tag{20}$$

Definition 5: Let $\mathring{\Theta}^i_t$, $t = 1, \ldots, T$, denote the posterior probability of Ξ^i_t given the common information $(Z_{1:t}, \mathbf{U}_{1:t-1})$, i.e., for any Borel subset E^i of $\Delta(\mathcal{X}^i)$,

$$\mathring{\Theta}^i_t(E^i) := \mathbb{P}(\Xi^i_t \in E^i \mid Z_{1:t}, \mathbf{U}_{1:t-1}).$$

Now, by following the exact same argument as in Steps 2 and 3 for the full observation model, we get that Propositions 5 and 6 and Theorems 1 and 2 are also true for the partial observation model if we replace Π_t and Θ^i_t by $\mathring{\Pi}_t$ and $\mathring{\Theta}^i_t$, respectively.

V. EXTENSION TO INFINITE HORIZON

In this section, we extend the result of structural result of Theorem 1 and the dynamic programming decomposition of Theorem 2 to a time-homogeneous system that runs for an infinite horizon under the discounted cost optimality criterion.

In the model of Section II-A, assume that the plant function f^i_t, $i = 0, \ldots, n$, and the cost function c_t are time-invariant and are denoted by f^i and c, respectively. Furthermore, in the partial observation model assume that the observation function ℓ^i_t, $i = 1, \ldots, n$ are time-invariant and are denoted by ℓ^i. Such a system is called a *time-homogeneous system*.

Assume that the system runs indefinitely. Define the performance of a control strategy $\mathbf{g} := (\mathbf{g}_1, \mathbf{g}_2, \ldots)$ as

$$J_\beta(\mathbf{g}) := \lim_{T \to \infty} \mathbb{E}\Big[\sum_{t=1}^T \beta^{t-1} c(Z_t, \mathbf{X}_t, \mathbf{U}_t)\Big], \tag{21}$$

where $\beta \in (0, 1)$ is called the discount factor.

We are interested in the following optimization problem.

Problem 2: Given a discount factor β, the distributions P_Z, $P_{X^i|Z}$, P_{W^i}, $P_{\tilde{W}^i}$ of the initial shared state, initial local state, plant disturbance of subsystem i, and observation noise of subsystem i (for the partial observation model), $i = 1, \ldots, n$, and the cost functions c, find a control strategy $\mathbf{g}$ that minimizes the expected discounted cost given by (21).

Since the sufficient statistic in Theorem 2 takes value in a time-invariant space, the results of the finite horizon system extend to infinite horizon in the usual manner. Proposition 2 remains valid

for an infinite horizon system as well. Consequently, so do the structural results of Proposition 1. Therefore, we can use the approach of [26] to obtain the infinite horizon version of the dynamic program of Proposition 6. Using Lemma 8, the dynamic program simplifies as follows:

Theorem 3: There exists an optimal control strategy that is time homogeneous. An optimal choice of the partially evaluated control strategy $\tilde{d}$ of $\tilde{g}$ is given by solution of the following fixed point equation[3] for the full observation model:

$$V(z, \boldsymbol{\theta}) = \min_{\tilde{\mathbf{d}}} \mathbb{E}\left[c(Z_t, X_t, U_t) + \beta V(Z_{t+1}, \boldsymbol{\Theta}_{t+1}) \,\Big|\, Z_t = z, \boldsymbol{\Theta}_t = \boldsymbol{\theta}, \tilde{\mathbf{D}}_t = \tilde{\mathbf{d}}\right] \tag{22}$$

and, by replacing Θ_t^i by $\mathring{\Theta}_t^i$ in the above equation for the partial observation model. (The above equation is time homogeneous; we are using time t for ease of notation.)

VI. AN EXAMPLE: MULTIACCESS BROADCAST

In this section, we reconsider the multiaccess broadcast system described in Section II-B and show how the results of this paper provide new insights for that system.

A. The model

Recall that a two-user multiaccess system consists is a special case of the full observation model with $\mathcal{X}^i = \mathcal{U}^i = \mathcal{W}_t^i = \{0, 1\}$, $i = 1, 2$, and $\mathcal{Z} = \emptyset$. The state dynamics of user 1 are given by: $X_{t+1}^1 = \max(X_t^1 + U_t^1 \cdot (1 - U_t^2) + W_t^1, 1)$. The dynamics of user 2 are symmetric dual of the above. Each user chooses a transmission decision as $U_t^i = g_t^i(X_{1:t}^i, \mathbf{U}_{1:t-1})$ where only actions $U_t^i \leq X_t^i$ are feasible. The per unit reward function $c(\mathbf{x}, \mathbf{u}) = u^1 \oplus u^2$, where $\oplus$ means binary XOR. The objective is to maximize the total average reward over an infinite horizon given by

$$\bar{J}(\mathbf{g}) = \lim_{T \to \infty} \frac{1}{T} \mathbb{E}\left[\sum_{t=1}^{T} U_t^1 \oplus U_t^2\right] \tag{23}$$

which corresponds to maximizing the average throughput.

The case of symmetric arrivals ($p^1 = p^2$) was considered in [1], who found a lower bound on performance by finding the best *window protocol* strategies. An upper bound was for the symmetric case was computed numerically in [2] by considering a more informative information structure. The analytic lower bounds of [1] match the numerical upper bound of [2]; hence,

---

[3]Due to the discounting of future costs, (22) has a fixed point that is unique.

the strategy proposed in [1] is optimal. A dynamic programming decomposition for the general model was presented in [28].

The multiaccess broadcast system corresponds to the full observation model. Therefore, the results of this paper[4] provide a structure of optimal transmission policies and a dynamic programming decomposition. For the symmetric arrival case ($p^1 = p^2$), we solve the corresponding dynamic program in closed form, and give an analytic derivation of the optimal strategy.

B. Structure of optimal transmission policies and dynamic programming decomposition

Since $Z_t = \emptyset$, the information state $\boldsymbol{\Theta}_t = (\Theta_t^1, \ldots, \Theta_t^n)$ of Definition 2 simplifies to $\Theta_t^i(x) = \mathbb{P}^{\tilde{\mathbf{g}}}(X_t^i = x \mid \mathbf{U}_{1:t-1})$. Theorem 1 implies that there is no loss of optimality in restricting attention to control strategies of the form $U_t^i = g_t^i(X_t^i, \boldsymbol{\Theta}_t)$ and Theorem 2 gives the corresponding dynamic program to find the optimal transmission strategies.

To succinctly describe the dynamic program, we simplify the notation as follows:

1) The functional map $\tilde{d}_t^i$ from $\mathcal{X}^i$ to $\mathcal{U}^i$ is completely specified by $\tilde{d}_t^i(1)$ because $\tilde{d}_t^i(0)$ must be zero as $u_t^i = \tilde{d}_t^i(x_t^i) \leq x_t^i$ and $\mathcal{X}^i = \mathcal{U}^i = \{0, 1\}$. We denote $\tilde{d}_t^i(1)$ by $S_t^i \in \{0, 1\}$. Then, $u_t^i = x_t^i \cdot s_t^i$.

2) Since Θ_t^i is a probability distribution of a binary valued random variable, it is completely specified by its component $\Theta_t^i(1)$, which we denote by Q_t^i.

To present the update of $\mathbf{Q}_t$, we define the following operators.

Definition 6: Let A_i, $i = 1, 2$, be an operator from $[0, 1]$ to $[0, 1]$ defined for any $q \in [0, 1]$ as $A_i q = 1 - (1 - p^i)(1 - q)$ where p^i is the arrival rate at the queue i. Then, $A_i^n q = 1 - (1 - p^i)^n(1 - q)$, and for any $q \in (0, 1)$, $A_i^n q < A_i^{n+1} q$.

Lemma 7 shows that the information state $\mathbf{q}_t = (q_t^1, q_t^2)$ updates according to a non-linear

---

[4] Although in Section V, we only considered the infinite horizon discounted cost criterion, the same argument also works for the average reward per unit time.

filter $F(\mathbf{q}_t, \mathbf{u}_t, \mathbf{s}_t)$ where

$$F((q^1, q^2), \mathbf{u}, \mathbf{s}) = \begin{cases} (A_1 q^1, A_2 q^2), & \text{if } \mathbf{s} = (0, 0) \\ (p^1, A_2 q^2), & \text{if } \mathbf{s} = (1, 0) \\ (A_1 q^1, p^2), & \text{if } \mathbf{s} = (0, 1) \\ (1, 1), & \text{if } \mathbf{s} = (1, 1) \text{ and } \mathbf{u} = (1, 1) \\ (p^1, p^2), & \text{if } \mathbf{s} = (1, 1) \text{ and } \mathbf{u} \neq (1, 1). \end{cases} \tag{24}$$

Substituting this update function in the infinite horizon average reward per unit time version of the dynamic program of Theorem 2, we get

Proposition 14: For the two-user multiaccess broadcast system, there is no loss in optimality in restricting attention to time-homogeneous transmission strategies of the form

$$U_t^i = \tilde{g}_t^i(X_t^i, \mathbf{Q}_t) = S_t^i(\mathbf{Q}_t) \cdot X_t^i.$$

An optimal strategy of such form is given by the solution of the following fixed point equation:

$$v(q^1, q^2) + J^* = \max\{v_{10}(q^1, q^2), v_{01}(q^1, q^2), v_{11}(q^1, q^2)\} \tag{25}$$

where J^* denotes the average reward per unit time, $v(q^1, q^2)$ is the relative value function at (q^1, q^2) and $v_{ij}(q^1, q^2)$ is the relative value-action function at (q^1, q^2) when (s^1, s^2) is chosen to be (i, j), $i, j \in \{0, 1\}$, *i.e.,*

$$v_{10}(q^1, q^2) = q^1 + v(p^1, A_2 q^2),$$

$$v_{01}(q^1, q^2) = q^2 + v(A_1 q^1, p^2),$$

$$v_{11}(q^1, q^2) = q^1 + q^2 - 2q^1 q^2 + q^1 q^2 v(1, 1) + (1 - q^1 q^2) v(p^1, p^2).$$

Some remarks:

1) We ruled out the action $(s^1, s^2) = (0, 0)$ because it is dominated by the action $(s^1, s^2) = (1, 0)$.

2) The information state $\mathbf{q}$ takes values in the uncountable set $[0, 1]^2$. However, the form of the non-linear filter F (24) implies that the reachable set of $\mathbf{q}$ is countable and is given by

$$\mathcal{R} = \{(1, 1), (1, p^2), (p^1, 1), (p^1, p^2)\} \cup \{(p^1, A_2^n p^2) : n \in \mathbb{N}\} \cup \{(A_1^n p_1, p^2) : n \in \mathbb{N}\}$$

Thus, we need to solve the dynamic program of Proposition 14 only for $\mathbf{q} \in \mathcal{R}$.

3) A similar dynamic programming decomposition for the two-user multiaccess broadcast channel was derived in [28], but [28] did not completely exploit the information structure of the system. In particular, [28] *a priori* restricted attention to transmission strategies of the form (9) while we show that such a restriction is without loss of optimality. Furthermore, the dynamic program in [28] is similar to that of Proposition 6 while we use a simpler form of the dynamic program (Theorem 2). As shown above, the reachable set of the information state is countable for this simpler form of the dynamic program while such a simplification was not possible for the dynamic program in [28].

C. The symmetric arrival case

Assume that both users have symmetric arrivals, *i.e.*, $p^1 = p^2$. Then the transformation A_1 is the same as A_2, and we denote both by A. Since the system is symmetric for both users, we have that for any q^1, q^2, the relative value functions $v(q^1, q^2)$ and $v(q^2, q^1)$ are the same. Therefore $v_{ij}(q^1, q^2) = v_{ji}(q^2, q^1)$, and consequently, the optimal coordination policy is also symmetric, *i.e.*, $h(q^1, q^2) = h(q^2, q^1)$.

Using this symmetry, we find a closed form solution of the dynamic program. To describe the solution, we first consider the following polynomial and some of its properties. Let $\varphi_n(x) = 1 + (1 - x)^2 - (3 + x)(1 - x)^{n+1}$. Note that

1) $\varphi_n(0) = -1$ and $\varphi_n(1) = 1$. Thus, φ_n has a root α_n that lies in the interval $[0, 1]$.

2) $\varphi_{n+1}(x) = (1 - x)\varphi_n(x) + x(1 + (1 - x)^2)$. Thus, $\varphi_{n+1}(\alpha_n)$ is positive. Recall that $\varphi_n(0) = -1$. Thus, α_{n+1} lies in the interval $[0, \alpha_n]$. Hence the sequence $\{\alpha_n\}$ is decreasing.

Let τ denote the root of $x = (1 - x)^2$. Then, $\tau \approx 0.38196 > \alpha_1 \approx 0.34727$.

Theorem 4: For the symmetric arrival case, $p^1 = p^2 = p$, the optimal solution J^* to the dynamic program of Proposition 14 is given by

$$J^* = \begin{cases} (1 - (1 - p)^2), & \text{if } p \geq \alpha_1, \\ p(1 - (2p^2 - 1))/(1 + p^2 + p^3), & \text{otherwise.} \end{cases} \tag{26}$$

The corresponding optimal strategy $h^*(q^1, q^2)$, $(q^1, q^2) \in \mathcal{R}$, is given by

1) For $p \geq \tau$,

$$h^*(q^1, q^2) = \begin{cases} (1, 0), & \text{if } q^1 > q^2, \\ (0, 1), & \text{if } q^1 < q^2, \\ (1, 0) \text{ or } (0, 1), & \text{if } q^1 = q^2. \end{cases}$$

2) For $p < \tau$, let $n \in \mathbb{N}$ be such that $\alpha_{n+1} < p \leq \alpha_n$. Then,

$$h^*(q^1, q^2) = \begin{cases} (1,1), & \text{if } q^1 \leq A^n p \text{ and } q^2 \leq A^n p, \\ (1,0), & \text{if } q^1 > \max(A^n p, q^2), \\ (0,1), & \text{if } q^2 > \max(A^n p, q^1), \\ (1,0) \text{ or } (0,1), & \text{if } q^1 = q^2 = 1. \end{cases}$$

The proof is presented in Appendix **G**.

Although the optimal policy looks complicated with different behavior depending on the value of p, it has only two modes of operation. When $p \geq \tau$, the set of states $\{(p, Ap), (Ap, p)\}$ is absorbing and forms a recurrence class in $\mathcal{R}$. Within this recurrence class, the optimal policy is a round-robin policy. When $p < \tau$, the set of states $\{(1,1), (p, Ap), (Ap, p), (p, p)\}$ is absorbing and forms a recurrence class in $\mathcal{R}$. Within this recurrence class, the optimal policy is identical for all $p < \tau$. The system starts with $(q^1, q^2) = (p, p)$ and chooses $(s^1, s^2) = (1, 1)$, which means that each user transmits if it has a packet. If no collision occurs, then the next state remains (p, p). If a collision occurs, $(q^1, q^2) = (1, 1)$ and both users know that both of them have a packet. So, they simply empty their buffer one by one, say first $(s^1, s^2) = (1, 0)$, and then $(s^1, s^2) = (0, 1)$, and go back to "transmit if you have a packet" action: $(s^1, s^2) = (1, 1)$. This policy is identical to the optimal window protocol proposed in [1]. Unlike [1], who showed that this strategy is the best transmission strategy when restricted to window protocols, we have shown that this strategy is the best strategy over the class of all transmission protocols.

VII. DISCUSSION AND CONCLUSION

Systems with control sharing information structure arise in a variety of communication applications. In this paper, we presented a three step approach to identify sufficient statistic and dynamic programming decomposition for coupled subsystems with control sharing.

The general decentralized control system with control sharing does not admit a tractable dynamic programming decomposition. Our solution approach works because the subsystems are coupled *only* through control actions $\mathbf{U}_t$ and shared state Z_t, *but not through local states* $\mathbf{X}_t$. In particular, if the system dynamics were of the form

$$X_{t+1}^i = f_t^i(Z_t, \mathbf{X}_t, \mathbf{U}_t, W_t^i) \tag{27}$$

instead of (2), then Propositions 2 and 11 will fail, and consequently, Step 1 of our approach would not simplify the control strategies.

In addition, the final sufficient statistics Θ_t and $\mathring{\Theta}_t$ derived in Step 3 are simpler than the general sufficient statistics Π_t and $\mathring{\Pi}_t$, which are based on [22], derived in Step 2. In particular, $\Pi_t \in \Delta(\mathcal{X}^1 \times \cdots \times \mathcal{X}^n)$, so its size increases exponentially with the number of subsystems, while $\mathring{\Pi}_t \in \Delta(\mathcal{X}^1) \times \cdots \times \Delta(\mathcal{X}^n)$, so its size increases linearly with the number of subsystems. This additional simplification is also a consequence of the specific form of system dynamics and would fail if the system dynamics were of the form (27).

In itself, it is not surprising that a simpler dynamical model makes the system easier to design. However, it is important to understand why *this particular* simplification dynamical model works; such an understanding will allow for similar simplifications for general non-classical information structures as well.

The system dynamics given by (2) do not remove the incentive to signal. In particular, control station i at time $t+1$ does not know all observations of control station j at time t. Hence, control station j has an incentive to signal its local observation to control station i through its action U_t^j. Thus, the model is not partially nested [14] (or quasi-classical [39]). Even after taking the conditional independence results of Propositions 2 and 11 into account, the signaling incentive is still present due to the cost coupling. Knowing the local state X_t^j of subsystem j will help control station i to improve its choice of action X_t^i in order to minimize the expected cost to go $\mathbb{E}[\sum_{s=t}^T c_s(Z_s, \mathbf{X}_s, \mathbf{U}_s)]$. Thus, the model is not stochastically nested [16] (or P-quasi-classical [39]).

We may think of the system dynamics of the form (2) as a sufficient condition to obtain a time-invariant sufficient statistic for the local information at each control stations (Propositions 1 and 10). Once such a sufficient statistic is identified, the model reduces to a partial history sharing information structure with *local information taking values in a time-invariant space*. Thereafter, one can use the results of [22] obtain a sufficient statistic of the common information at all control stations.

Finding such sufficient conditions (to extend the applicability of a specific solution technique to more general models) is a recurring theme in decentralized control. A similar approach has been used in [10] to generalize the solution approach of [8] to two controller teams where at least one controller has finite memory; in [16] to generalize the solution approach of [14] to stochastically

nested information structures; in [40] to generalize the solution approach of [28] to broadcast information structures; and in [39] to generalize the solution of classical and quasiclassical information structures to P-classical and P-quasiclassical information structures. The model and results of this paper present such a sufficient condition to extend the results of [22] to control sharing information structure.

ACKNOWLEDGMENT

REFERENCES

[1] M. G. Hluchyj and R. G. Gallager, "Multiacces of a slotted channel by finitely many users," in *Proceedings of National Telecommunication Conference*, 1981, pp. D.4.2.1–D.4.2.7.

[2] J. M. Ooi and G. W. Wornell, "Decentralized control of a multiple access broadcast channel: performance bounds," in *Proceedings of the 35th IEEE Conference on Decision and Control*, Kobe, Japan, 1996, pp. 293–298.

[3] B. Hajek, K. Mitzel, and S. Yang, "Paging and registration in cellular networks: Jointly optimal policies and an iterative algorithm," *IEEE Trans. Inf. Theory*, vol. 64, pp. 608–622, Feb. 2008.

[4] J. C. Walrand and P. Varaiya, "Optimal causal coding-decoding problems," *IEEE Trans. Inf. Theory*, vol. 29, no. 6, pp. 814–820, Nov. 1983.

[5] V. V. Veeravalli, T. Başar, and H. V. Poor, "Decentralized sequential detection with a fusion center performingthe sequential test," *IEEE Transactions on Information Theory*, vol. 39, no. 2, pp. 433–442, 1993.

[6] H. S. Witsenhausen, "Separation of estimation and control for discrete time systems," *Proc. IEEE*, vol. 59, no. 11, pp. 1557–1566, Nov. 1971.

[7] Y.-C. Ho, "Team decision theory and information structures," *Proc. IEEE*, vol. 68, no. 6, pp. 644–654, 1980.

[8] H. S. Witsenhausen, "A standard form for sequential stochastic control," *Mathematical Systems Theory*, vol. 7, no. 1, pp. 5–11, 1973.

[9] A. Nayyar, "Sequential decision making in decentralized systems," Ph.D. dissertation, University of Michigan, Ann Arbor, MI, 2011.

[10] A. Mahajan, "Sequential decomposition of sequential dynamic teams: applications to real-time communication and networked control systems," Ph.D. dissertation, University of Michigan, Ann Arbor, MI, Sep. 2008.

[11] S. Yüksel and T. Başar, *Stochastic Networked Control Systems: Stabilization and Optimization under Information Constraints*. (under preparation), 2012.

[12] R. Radner, "Team decision problems," *Annals of Mathmatical Statistics*, vol. 33, pp. 857–881, 1962.

[13] J. Krainak, J. Speyer, and S. Marcus, "Static team problems–part I: Sufficient conditions and the exponential cost criterion," *IEEE Trans. Autom. Control*, vol. 27, no. 4, pp. 839–848, Aug. 1982.

[14] Y.-C. Ho and K.-C. Chu, "Team decision theory and information structures in optimal control problems–Part I," *IEEE Trans. Autom. Control*, vol. 17, no. 1, pp. 15–22, 1972.

[15] A. Gattami, "Generalized linear quadratic control," Ph.D. dissertation, Lund University, May 2007.

[16] S. Yüksel, "Stochastic nestedness and the belief sharing information pattern," *IEEE Trans. Autom. Control*, pp. 2773–2786, Dec. 2009.

[17] B. Bamieh and P. Voulgaris, "A convex characterization of distributed control problems in spatially invariant systems with communication constraints," *Systems and Control Letters*, vol. 54, no. 6, pp. 575–583, 2005.

[18] M. Roktowitz and S. Lall, "A characterization of convex problems in decentralized control," *IEEE Trans. Autom. Control*, vol. 51, no. 2, pp. 274–286, 2006.

[19] A. Rantzer, "Linear quadratic team theory revisited," in *Proceedings of the American Control Confernece*, 2006, pp. 1637–1641.

[20] M. Aicardi, F. Davoli, and R. Minciardi, "Decentralized optimal control of Markov chains with a common past information set," *IEEE Trans. Autom. Control*, vol. 32, no. 11, pp. 1028–1031, 1987.

[21] P. Varaiya and J. Walrand, "On delayed sharing patterns," *IEEE Trans. Autom. Control*, vol. 23, no. 3, pp. 443–445, 1978.

[22] A. Nayyar, A. Mahajan, and D. Teneketzis, "Optimal control strategies in delayed sharing information structures," *IEEE Trans. Autom. Control*, vol. 56, no. 7, pp. 1606–1620, Jul. 2011.

[23] J.-M. Bismut, "An example of interaction between information and control: The transparency of a game," *IEEE Trans. Autom. Control*, vol. 18, no. 5, pp. 518–522, Oct. 1972.

[24] N. Sandell and M. Athans, "Solution of some nonclassical lqg stochastic decision problems," *IEEE Trans. Autom. Control*, vol. 19, pp. 108–116, 1974.

[25] J. M. Ooi, S. M. Verbout, J. T. Ludwig, and G. W. Wornell, "A separation theorem for periodic sharing information patterns in decentralized control," *IEEE Trans. Autom. Control*, vol. 42, no. 11, pp. 1546–1550, Nov. 1997.

[26] A. Nayyar, A. Mahajan, and D. Teneketzis, "Dynamic programming for multi-controller stochastic control with partial information sharing: A common-information approach," 2011, submitted to IEEE Trans. Autom. Control.

[27] H. S. Witsenhausen, "On the structure of real-time source coders," *Bell System Technical Journal*, vol. 58, no. 6, pp. 1437–1451, July-August 1979.

[28] A. Mahajan, A. Nayyar, and D. Teneketzis, "Identifying tractable decentralized control problems on the basis of information structure," in *Proc. 46th Annual Allerton Conf. Communication, Control, and Computing*, Monticello, IL, Sep. 2008, pp. 1440–1449.

[29] F. C. Schoute, "Decentralized control in packet switched satellite communication," *IEEE Trans. Autom. Control*, vol. AC-23, no. 2, pp. 362–271, Apr. 1976.

[30] P. Varaiya and J. Walrand, "Decentralized control in packet switched satellite communication," *IEEE Trans. Autom. Control*, vol. AC-24, no. 5, pp. 794–796, Oct. 1979.

[31] E. A. Hansen, D. S. Bernstein, and S. Zilberstein, "Dynamic programming for partially observable stochastic games," in *Proceedings of the 19th national conference on artificial intelligence (AAAI)*, San Jose, CA, Jul. 2004, pp. 709–715.

[32] D. Szer, F. Charpillet, and S. Zilberstein, "MAA*: A heuristic search algorithm for solving decentralized POMDPs," in *Proceedings of the 21st Conference on Uncertainty in Artificial Intelligence (UAI)*, 2005.

[33] D. Szer and F. Charpillet, "An optimal best-first search algorithm for solving infinite horizon DEC-POMDPs," in *Machine Learning*, ser. Lecture Notes in Computer Science, J. a. Gama, R. Camacho, P. Brazdil, A. Jorge, and L. Torgo, Eds. Berlin, Heidelberg: Springer Berlin / Heidelberg, 2005, vol. 3720, ch. 38, pp. 389–399.

[34] S. Seuken and S. Zilberstein, "Memory-bounded dynamic programming for DEC-POMDPs," in *Proceedings of the 20th international joint conference on Artifical intelligence (IJCAI)*, San Francisco, CA, USA, 2007, pp. 2009–2015.

[35] J. S. Dibangoye, A.-I. Mouaddib, and B. Chaib-draa, "Incremental pruning heuristic for solving DEC-POMDPs," in *Proceedings of the Workshop on Multi-Agent Sequential Decision Making in Uncertain Domains (MSDM)*, May 2008.

[36] O. Hernández-Lerma and J. Lasserre, *Discrete-Time Markov Control Processes.* Springer-Verlag, 1996.

[37] P. Whittle, *Optimization Over Time*, ser. Wiley Series in Probability and Mathematical Statistics. John Wiley and Sons, 1983, vol. 2.

[38] C. Striebel, "Sufficient statistics in the optimal control of stochastic systems," *Journal of Mathematical Analysis and Applications*, vol. 12, pp. 576–592, 1965.

[39] A. Mahajan and S. Yksel, "Measure and cost dependent properties of information structures," in *Proc. American Control Conf. (ACC)*, Baltimore, MD, Jul. 2010, pp. 6397–6402.

[40] J. Wu and S. Lall, "A dynamic programming algorithm for decentralized markov decision processes with a broadcast structure," in *Proceedings of the 49th IEEE Conference on Decision and Control*, 2010, pp. 6143–6148.

APPENDIX A

PROOF OF PROPOSITION 2

For simplicity of notation, we use $\mathbb{P}(z_{1:t}, \mathbf{x}_{1:t}, \mathbf{u}_{1:t})$ to denote $\mathbb{P}(Z_{1:t} = z_{1:t}, \mathbf{X}_{1:t} = \mathbf{x}_{1:t}, \mathbf{U}_{1:t} = \mathbf{u}_{1:t})$ and a similar notation for conditional probability. Define:

- $\alpha_t^i \coloneqq \mathbb{P}(u_t^i \,|\, z_{1:t}, x_{1:t}^i, \mathbf{u}_{1:t-1})$, $\beta_t^i \coloneqq \mathbb{P}(x_t^i \,|\, z_{t-1}, x_{t-1}^i, \mathbf{u}_{t-1})$, $\gamma_t^i \coloneqq \mathbb{P}(z_t \,|\, z_{t-1}, \mathbf{u}_{t-1})$; and
- $A_t^i \coloneqq \prod_{s=1}^t \alpha_s^i$, $B_t^i \coloneqq \prod_{s=1}^t \beta_s^i$, $\Gamma_t \coloneqq \prod_{s=1}^t \gamma_s$.

From law of total probability it follows that: $\mathbb{P}(z_{1:t}, \mathbf{x}_{1:t}, \mathbf{u}_{1:t}) = \left(\prod_{i=1}^n A_t^i B_t^i \right) \Gamma_t$. Summing over all realizations of $\mathbf{x}_{1:t}$ and observing that A_t^i and B_t^i depends only on $(z_{1:t}, x_{1:t}^i, \mathbf{u}_{1:t})$, we get

$$\mathbb{P}(z_{1:t}, \mathbf{u}_{1:t}) = \sum_{x_{1:t}^1} \sum_{x_{1:t}^2} \cdots \sum_{x_{1:t}^n} \left(\prod_{i=1}^n A_t^i B_t^i \right) \Gamma_t = \left(\prod_{i=1}^n \left(\sum_{x_{1:t}^i} A_t^i B_t^i \right) \right) \Gamma_t.$$

Thus, using Bayes rule we get

$$\mathbb{P}(\mathbf{x}_{1:t} \,|\, z_{1:t}, \mathbf{u}_{1:t}) = \prod_{i=1}^n \frac{A_t^i B_t^i}{\left(\sum_{x_{1:t}^i} A_t^i B_t^i \right)} \tag{28}$$

Summing both sides over $x_{1:t}^i$, $i \neq j$, we get

$$\mathbb{P}(x_{1:t}^j \,|\, z_{1:t}, \mathbf{u}_{1:t}) = \frac{A_t^j B_t^j}{\left(\sum_{x_{1:t}^j} A_t^j B_t^j \right)} \tag{29}$$

The result follows from combining (28) and (29).

APPENDIX B

PROOF OF LEMMA 3

For ease of notation, we use $\mathbb{P}(\tilde{r}_{t+1}^i \mid r_{1:t}^i, u_{1:t}^i)$ to denote $\mathbb{P}(R_{t+1}^i = \tilde{r}_{t+1}^i \mid R_{1:t}^i = r_{1:t}^i, U_{1:t}^i = u_{1:t}^i)$ and a similar notation for other probability statements. Consider

$$\mathbb{P}(\tilde{r}_{t+1}^i \mid r_{1:t}^i, u_{1:t}^i) = \mathbb{P}(\tilde{x}_{t+1}^i \mid x_t^i, \tilde{z}_t, \tilde{\mathbf{u}}_t) \cdot \mathbb{P}(\tilde{z}_{t+1} \mid \tilde{z}_t, \tilde{\mathbf{u}}_t) \cdot \mathbb{1}[\tilde{\mathbf{u}}_{1:t-1} = \mathbf{u}_{1:t-1}] \cdot \mathbb{1}[\tilde{u}_t^i = u_t^i]$$
$$\cdot \mathbb{1}[\tilde{z}_{1:t} = z_{1:t}] \cdot \mathbb{P}(\tilde{\mathbf{u}}_t^{-i} \mid x_{1:t}^i, z_{1:t}, \mathbf{u}_{1:t-1}, u_t^i) \quad (30)$$

Simplify the last term of (30) as follows:

$$\mathbb{P}(\tilde{\mathbf{u}}_t^{-i} \mid x_{1:t}^i, z_{1:t}, \mathbf{u}_{1:t-1}, u_t^i) \overset{(a)}{=} \mathbb{P}(\tilde{\mathbf{u}}_t^{-i} \mid x_{1:t}^i, z_{1:t}, \mathbf{u}_{1:t-1})$$
$$= \sum_{\mathbf{x}_{1:t}^{-i}} \mathbb{P}(\tilde{\mathbf{u}}_t^{-i} \mid \mathbf{x}_{1:t}^{-i}, z_{1:t}, \mathbf{u}_{1:t-1}) \cdot \mathbb{P}(\mathbf{x}_{1:t}^{-i} \mid x_{1:t}^i, z_{1:t}, \mathbf{u}_{1:t-1})$$
$$\overset{(b)}{=} \sum_{\mathbf{x}_{1:t}^{-i}} \mathbb{P}(\tilde{\mathbf{u}}_t^{-i} \mid \mathbf{x}_{1:t}^{-i}, z_{1:t}, \mathbf{u}_{1:t-1}) \cdot \mathbb{P}(\mathbf{x}_{1:t}^{-i} \mid z_{1:t}, \mathbf{u}_{1:t-1}) = \mathbb{P}(\tilde{\mathbf{u}}_t^{-i} \mid z_{1:t}, \mathbf{u}_{1:t-1}) \quad (31)$$

where (a) is true because u_t^i is determined by $x_{1:t}^i$ $z_{1:t}$ and $\mathbf{u}_{1:t-1}$ and (b) follows from Proposition 2. Substituting (31) in (30), we get

$$\mathbb{P}(\tilde{r}_{t+1}^i \mid r_{1:t}^i, u_{1:t}^i) = \mathbb{P}(\tilde{x}_{t+1}^i \mid x_t^i, \tilde{z}_t, \tilde{\mathbf{u}}_t) \cdot \mathbb{P}(\tilde{z}_{t+1} \mid \tilde{z}_t, \tilde{\mathbf{u}}_t) \cdot \mathbb{1}[\tilde{\mathbf{u}}_{1:t-1} = \mathbf{u}_{1:t-1}]$$
$$\cdot \mathbb{1}[\tilde{z}_{1:t} = z_{1:t}] \cdot \mathbb{1}[\tilde{u}_t^i = u_t^i] \cdot \mathbb{P}(\tilde{\mathbf{u}}_t^{-i} \mid z_{1:t}, \mathbf{u}_{1:t-1})$$
$$= \mathbb{P}(\tilde{x}_{t+1}^i, \tilde{z}_{1:t+1}, \tilde{\mathbf{u}}_{1:t} \mid x_t^i, u_t^i, z_{1:t}, \mathbf{u}_{1:t-1}) = \mathbb{P}(\tilde{r}_{t+1}^i \mid r_t^i, u_t^i) \quad (32)$$

This completes the proof of part 1) of the Lemma.

To prove part 2), it is sufficient to show that $\mathbb{P}(\tilde{z}_t, \tilde{\mathbf{x}}_t, \tilde{\mathbf{u}}_t \mid r_{1:t}^i, u_{1:t}^i) = \mathbb{P}(\tilde{z}_t, \tilde{\mathbf{x}}_t, \tilde{\mathbf{u}}_t \mid r_t^i, u_t^i)$. Consider

$$\mathbb{P}(\tilde{z}_t, \tilde{\mathbf{x}}_t, \tilde{\mathbf{u}}_t \mid r_{1:t}^i, u_{1:t}^i) = \mathbb{1}[(\tilde{z}_t, \tilde{x}_t^i, \tilde{u}_t^i) = (z_t, x_t^i, u_t^i)] \cdot \mathbb{P}(\tilde{\mathbf{x}}_t^{-i}, \tilde{\mathbf{u}}_t^{-i} \mid x_{1:t}^i, u_t^i, z_t, \mathbf{u}_{1:t-1})$$
$$\overset{(c)}{=} \mathbb{1}[(\tilde{z}_t, \tilde{x}_t^i, \tilde{u}_t^i) = (z_t, x_t^i, u_t^i)] \cdot \mathbb{P}(\tilde{\mathbf{x}}_t^{-i}, \tilde{\mathbf{u}}_t^{-i} \mid z_{1:t}, \mathbf{u}_{1:t-1})$$
$$= \mathbb{P}(\tilde{\mathbf{x}}_t, \tilde{\mathbf{u}}_t \mid r_t^i, u_t^i) \quad (33)$$

where (c) follows from an argument similar to (31).[5] This completes the proof of part 2) of the Lemma.

---

[5]Recall that $\mathbf{x}_t^{-i}$ denotes the vector $(x_t^1, \ldots, x_t^{i-1}, x_t^{i+1}, \ldots, x_t^n)$.

APPENDIX C

PROOF OF LEMMA 7

Consider the system for a particular realization $(z_{1:T}, \mathbf{x}_{1:T}, \mathbf{u}_{1:T}, \mathbf{d}_{1:T})$ of $(Z_{1:T}, \mathbf{X}_{1:T}, \mathbf{U}_{1:T}, \mathbf{D}_{1:T})$. For ease of notation, we use $\mathbb{P}(x_{t+1}^i \mid z_{1:t+1}, \mathbf{u}_{1:t}, \mathbf{d}_{1:t})$ to denote $\mathbb{P}(X_{t+1}^i = x_{t+1}^i \mid Z_{1:t+1} = z_{1:t+1}, \mathbf{U}_{1:t} = \mathbf{u}_{1:t}, \mathbf{D}_{1:t} = \mathbf{d}_{1:t})$. Define

$$A(x_{t+1}^i, \mathbf{x}_t, z_{1:t+1}, \mathbf{u}_{1:t}, \mathbf{d}_{1:t}) := \mathbb{P}(x_{t+1}^i, \mathbf{x}_t, z_{t+1}, \mathbf{u}_t \mid z_{1:t}, \mathbf{u}_{1:t-1}, \mathbf{d}_{1:t});$$

$$B(x_{t+1}^i, \mathbf{x}_t, z_{t+1}, z_t, \mathbf{d}_t, \boldsymbol{\theta}_t) := \mathbb{P}(x_{t+1}^i \mid x_t^i, z_t, \mathbf{u}_t) \cdot \mathbb{P}(z_{t+1} \mid \mathbf{x}_t, z_t, \mathbf{u}_t) \cdot \prod_{i=1}^{n} \theta_t^i(x_t^i).$$

The system dynamics and Proposition 2 implies that

$$A(x_{t+1}^i, \mathbf{x}_t, z_{1:t+1}, \mathbf{u}_{1:t}, \mathbf{d}_{1:t}) = B(x_{t+1}^i, \mathbf{x}_t, z_{t+1}, z_t, \mathbf{d}_t, \boldsymbol{\theta}_t) \, \mathbb{1}[u_t = d_t(\mathbf{x}_t)] \tag{34}$$

Consider component-i of the realization $\boldsymbol{\theta}_{t+1}$ of $\boldsymbol{\Theta}_{t+1}$.

$$\theta_{t+1}^i(x_{t+1}^i) = \mathbb{P}(x_{t+1}^i \mid z_{1:t+1}, \mathbf{u}_{1:t}, \mathbf{d}_{1:t}) = \sum_{\{\mathbf{x}_t : d_t(\mathbf{x}_t) = \mathbf{u}_t\}} \frac{A(x_{t+1}^i, \mathbf{x}_t, z_{1:t+1}, \mathbf{u}_{1:t}, \mathbf{d}_{1:t})}{\sum_{\tilde{x}_{t+1}^i} A(\tilde{x}_{t+1}^i, \mathbf{x}_t, z_{1:t+1}, \mathbf{u}_{1:t}, \mathbf{d}_{1:t})}$$

$$\overset{(a)}{=} \sum_{\{\mathbf{x}_t : d_t(\mathbf{x}_t) = \mathbf{u}_t\}} \frac{B(x_{t+1}^i, \mathbf{x}_t, z_{t+1}, z_t, \mathbf{d}_t, \boldsymbol{\theta}_t)}{\sum_{\tilde{x}_{t+1}^i} B(\tilde{x}_{t+1}^i, \mathbf{x}_t, z_{t+1}, z_t, \mathbf{d}_t, \boldsymbol{\theta}_t)} =: F_t^i(\boldsymbol{\theta}_t, z_{t+1}, \mathbf{u}_t, \mathbf{d}_t)(x_{t+1}^i) \tag{35}$$

where (a) follows from (34). Combining (35) for all i, $i = 1, \ldots, n$, proves the Lemma.

APPENDIX D

PROOF OF PROPOSITION 11

The proof is similar to proof of Proposition 2. As before, for ease of notation, we use $\mathbb{P}(z_{1:t}, \mathbf{x}_{1:t}, \mathbf{y}_{1:t}, \mathbf{u}_{1:t})$ to denote $\mathbb{P}(Z_{1:t} = z_{1:t}, \mathbf{X}_{1:t} = \mathbf{x}_{1:t}, \mathbf{Y}_{1:t} = \mathbf{y}_{1:t}, \mathbf{U}_{1:t} = \mathbf{u}_{1:t})$. Define

- $\alpha_t^i = \mathbb{P}(u_t^i \mid z_{1:t}, y_{1:t}^i, \mathbf{u}_{1:t-1})$, $\beta_t^i = \mathbb{P}(x_t^i \mid z_{t-1}, x_{t-1}^i, \mathbf{u}_{1:t-1})$, $\gamma_t = \mathbb{P}(z_t \mid z_{t-1}, \mathbf{u}_t)$, $\delta_t^i = \mathbb{P}(y_t^i \mid x_t^i)$; and
- $A_t^i := \prod_{s=1}^{t} \alpha_s^i$, $B_t^i := \prod_{s=1}^{t} \beta_s^i$, $\Gamma_t := \prod_{s=1}^{t} \gamma_s$, $\Delta_t^i := \prod_{s=1}^{t} \delta_s^i$.

From the law of total probability, it follows that $\mathbb{P}(z_{1:t}, \mathbf{x}_{1:t}, \mathbf{y}_{1:t}, \mathbf{u}_{1:t}) = big(\prod_{i=1}^{n} A_t^i B_t^i \Delta_t^i)\Gamma_t^i$. Sum over the realizations of $\mathbf{y}_{1:t}$ and observe that A_t^i and Δ_t^i depend on $\mathbf{y}_{1:t}$ only through $y_{1:t}^i$. This gives,

$$\mathbb{P}(z_{1:t}, \mathbf{x}_{1:t}, \mathbf{u}_{1:t}) = \sum_{y_{1:t}^1} \sum_{y_{1:t}^2} \cdots \sum_{y_{1:t}^n} \left(\prod_{i=1}^{n} A_t^i B_t^i \Delta_t^i \right) \Gamma_t = \left(\prod_{i=1}^{n} \left(\sum_{y_{1:t}^i} A_t^i \Delta_t^i \right) B_t^i \right) \Gamma_t$$

Now sum over $\mathbf{x}_{1:t}$ and observe that B_t^i and Δ_t^i depend on $\mathbf{x}_{1:t}$ only through $x_{1:t}^i$.

$$\mathbb{P}(z_{1:t}, \mathbf{u}_{1:t}) = \sum_{x_{1:t}^1} \sum_{x_{1:t}^2} \cdots \sum_{x_{1:t}^n} \left(\prod_{i=1}^{n} \left(\sum_{y_{1:t}^i} A_t^i \Delta_t^i \right) B_t^i \right) \Gamma_t$$

$$= \left(\prod_{i=1}^{n} \left(\sum_{x_{1:t}^i} B_t^i \left(\sum_{y_{1:t}^i} A_t^i \Delta_t^i \right) \right) \right) \Gamma_t$$

Thus, by Bayes rule, we get

$$\mathbb{P}(\mathbf{x}_{1:t} \mid z_{1:t}, \mathbf{u}_{1:t}) = \prod_{s=1}^{n} \frac{B_t^i \left(\sum_{y_{1:t}^i} A_t^i \Delta_t^i \right)}{\sum_{x_{1:t}^i} B_t^i \left(\sum_{y_{1:t}^i} A_t^i \Delta_t^i \right)} \tag{36}$$

Summing both sides over $x_{1:t}^i$, $i \neq j$, we get

$$\mathbb{P}(x_{1:t}^j \mid z_{1:t}, \mathbf{u}_{1:t}) = \frac{B_t^j \left(\sum_{y_{1:t}^j} A_t^j \Delta_t^j \right)}{\sum_{x_{1:t}^j} B_t^j \left(\sum_{y_{1:t}^j} A_t^j \Delta_t^j \right)} \tag{37}$$

The result follows from combining (36) and (37).

APPENDIX E

PROOF OF PROPOSITION 12

Consider

$$\mathbb{P}(\mathbf{\Xi}_{1:t} \in \mathbf{E}_{1:t} \mid z_{1:t}, \mathbf{u}_{1:t}) = \int_{\mathbf{E}_{1:t}} \mathrm{d}\,\mathbb{P}(\boldsymbol{\xi}_{1:t} \mid z_{1:t}, \mathbf{u}_{1:t})$$

From Proposition 11 and law of total probability, we get

$$\mathrm{d}\,\mathbb{P}(\boldsymbol{\xi}_{1:t} \mid z_{1:t}, \mathbf{u}_{1:t}) = \sum_{\mathbf{x}_{1:t}, \mathbf{y}_{1:t}} \left(\prod_{i=1}^{n} \mathrm{d}\,\mathbb{P}(\xi_t^i \mid y_{1:t}^i, z_{1:t}, \mathbf{u}_{1:t}) \cdot \mathbb{P}(y_{1:t}^i \mid x_{1:t}^i) \cdot \mathbb{P}(x_{1:t}^i \mid z_{1:t}, \mathbf{u}_{1:t}) \right)$$

$$= \prod_{i=1}^{n} \left(\sum_{\mathbf{x}_{1:t}, \mathbf{y}_{1:t}} \mathrm{d}\,\mathbb{P}(\xi_t^i \mid y_{1:t}^i, z_{1:t}, \mathbf{u}_{1:t}) \cdot \mathbb{P}(y_{1:t}^i \mid x_{1:t}^i) \cdot \mathbb{P}(x_{1:t}^i \mid z_{1:t}, \mathbf{u}_{1:t}) \right)$$

which completes the proof of the Proposition.

APPENDIX F

PROOF OF LEMMA 13

For ease of notation, we use $\mathrm{d}\,\mathbb{P}(\tilde{r}_{t+1}^i \mid r_{1:t}^i, u_{1:t-1}^i)$ to denote $\mathrm{d}\,\mathbb{P}(R_{t+1}^i = \tilde{r}_{t+1}^i \mid R_{1:t}^i = r_{1:t}^i, U_{1:t-1}^i = u_{1:t-1}^i)$ and a similar notation for other probability measures. Consider

$$\mathrm{d}\,\mathbb{P}(\tilde{r}_{t+1}^i \mid r_{1:t}^i, u_{1:t}^i) = \sum_{x_{t:t+1}^i, y_{t+1}^i} \mathbb{1}[\tilde{\xi}_{t+1}^i = \tilde{F}_t^i(\xi_t^i, y_{t+1}^i, \tilde{z}_t, \tilde{\mathbf{u}}_t)] \cdot \mathbb{P}(y_{t+1}^i \mid x_{t+1}^i) \cdot \mathbb{P}(x_{t+1}^i \mid x_t^i, \tilde{z}_t, \tilde{\mathbf{u}}_t)$$

$$\cdot \, \mathbb{P}(\tilde{z}_{t+1} \mid \tilde{z}_t, \tilde{\mathbf{u}}_t) \cdot \mathbb{1}[\tilde{z}_{1:t} = z_{1:t}] \cdot \mathbb{1}[\tilde{u}_t^i = u_t^i] \cdot \mathbb{1}[\tilde{\mathbf{u}}_{1:t-1} = \mathbf{u}_{1:t-1}]$$

$$\cdot \, \xi_t^i(x_t^i) \cdot \mathbb{P}(\mathbf{u}_t^{-i} \mid \xi_{1:t}^i, z_{1:t}, \mathbf{u}_{1:t-1}, u_t^i) \tag{38}$$

Simplify the last term of (38) as follows:

$$\mathbb{P}(\mathbf{u}_t^{-i} \mid \xi_{1:t}^i, z_{1:t}, \mathbf{u}_{1:t-1}, u_t^i)$$

$$= \sum_{\mathbf{x}_{1:t}^{-i}, \mathbf{y}_{1:t}^{-i}} \mathbb{P}(\mathbf{u}^{-i} \mid \mathbf{y}_{1:t}^{-i}, z_{1:t}, \mathbf{u}_{1:t-1}) \cdot \mathbb{P}(\mathbf{y}_{1:t}^{-i} \mid \mathbf{x}_{1:t}^{-i}) \cdot \mathbb{P}(\mathbf{x}_{1:t}^{-i} \mid z_{1:t}, \mathbf{u}_{1:t-1})$$

$$= \mathbb{P}(\mathbf{u}_t^{-i} \mid z_{1:t}, \mathbf{u}_{1:t-1}) \tag{39}$$

Substituting (39) in (38) and simplifying, we get part 1) of the Lemma:

$$\mathrm{d}\,\mathbb{P}(\tilde{r}_{t+1}^i \mid r_{1:t}^i, u_{1:t}^i) = \mathrm{d}\,\mathbb{P}(\tilde{r}_{t+1}^i \mid r_t^i, u_t^i) \tag{40}$$

The proof of part 2) is similar to (33).

APPENDIX G

PROOF OF THEOREM 4

We introduce a short hand notation that exploits the symmetry of the problem and the fact that the reachable set $\mathcal{R}$ is countable. Define

$$
\begin{aligned}
v^* &= v(1,1), & v^0 &= v(p,p), & v^n &= v(p, A^n p), \ n \in \mathbb{N}, & v^\infty &= v(p,1) \\
a^* &= v_{10}(1,1), & a^0 &= v_{10}(p,p), & a^n &= v_{10}(p, A^n p), \ n \in \mathbb{N}, & a^\infty &= v_{10}(p,1) \\
b^* &= v_{01}(1,1), & b^0 &= v_{01}(p,p), & b^n &= v_{01}(p, A^n p), \ n \in \mathbb{N}, & b^\infty &= v_{01}(p,1) \\
c^* &= v_{11}(1,1), & c^0 &= v_{11}(p,p), & c^n &= v_{11}(p, A^n p), \ n \in \mathbb{N}, & c^\infty &= v_{11}(p,1)
\end{aligned}
$$

Notice that $v_{10}(A^n p, p) = v_{01}(p, A^n p) = b^n$ and $v_{01}(A^n p, p) = v_{10}(p, A^n p) = a^n$.

With the above notation, the dynamic program of Proposition 14 can be written as

$$v^n + J^* = \max\{a^n, b^n, c^n\}, \quad n \in \{*, 0, 1, 2, \dots, \infty\} \tag{41}$$

where

$$a^* = 1 + v^\infty, \qquad b^* = 1 + v^\infty, \qquad c^* = v^* \tag{42a}$$

$$a^0 = p + v^1, \qquad b^0 = p + v^1, \qquad c^0 = 2p(1-p) + p^2 v^* + (1-p^2)v^0 \tag{42b}$$

$$a^n = p + v^{n+1}, \qquad b^n = A^n p + v^1, \qquad c^n = p + A^n p - 2p \cdot A^n p + p \cdot A^n p \cdot v^* \tag{42c}$$
$$+ (1 - p \cdot A^n p) v^0$$

$$a^\infty = p + v^\infty, \qquad b^\infty = 1 + v^1, \qquad c^\infty = 1 - p + p v^* + (1 - p) v^0 \tag{42d}$$

Lemma 15: A solution of the fixed point equations of (42) is given by the following:

1) For $p \in (\tau, 1]$

$$J^* = Ap, \quad v^* = 2 - Ap, \quad v^0 = p, \quad v^n = A^n p, \ n \in \mathbb{N}, \quad v^\infty = 1. \tag{43}$$

2) For $p \in (\alpha_1, \tau]$

$$J^* = Ap, \quad v^* = 2 - Ap, \quad v^0 = 1 - Ap, \quad v^n = A^n p, \ n \in \mathbb{N}, \quad v^\infty = 1. \tag{44}$$

3) For $p \in (\alpha_{m+1}, \alpha_m]$, $m \in \mathbb{N}$, define $\zeta(x) = 1 + x^2 + x^3$. Then,

$$J^* = p \left(1 - \frac{\varphi_0(p)}{\zeta(p)} \right), \quad v^* = 2 - J^*, \quad v^0 = 2 - p - \frac{1 + (1 - p)^2}{\zeta(p)}, \quad v^\infty = 1. \tag{45a}$$

and

$$v^n = \begin{cases} w^n, & \text{if } n \leq m, \\ A^n p, & \text{if } n > m; \end{cases} \quad n \in \mathbb{N}, \tag{45b}$$

where

$$w^n = (1 - p) A^{n-1} p \left(\frac{J^*}{p} - (1 - p) \right) + v^0$$

Note that $w^1 = J^*$.

The proof follows from elementary algebra. For completeness, we include the details for each case below.

Case 1: $p \in (\tau, 1]$

We show that the values of Lemma 15 satisfy the dynamic program of (41) and (42), by considering the four cases separately.

1) $a^* = b^* = 2$, $c^* = 2 - Ap$. Hence, either action $(0, 1)$ or $(1, 0)$ is optimal at state $(1, 1)$ and $v^* + J^* = a^* = b^* = 2$.

2) $a^0 = b^0 = p + Ap$, and $b^0 - c^0 = p^2(p - (1 - p)^2)$ which is positive for $p > \tau$. Recall that τ is the root of $x = (1 - x)^2$. Hence, either action $(0, 1)$ or $(1, 0)$ is optimal at state (p, p) and $v^0 + J^* = a^0 = b^0 = p + Ap$.

3) Consider $n \in \mathbb{N}$. $a^n = p + A^{n+1}p$ and $b^n = A^n p + Ap$. Thus, $b^n - a^n = p(1-p) \cdot A^{n-1}p \geq 0$. Moreover, $b^n - c^n = p^2[(3-p)A^n p - 1] > p^2[(3-p)p - 1]$ which is positive for $p \in (\tau, 1]$. Thus, the action $(0, 1)$ is optimal at state $(p, A^n p)$ (and by symmetry, the action $(1, 0)$ is optimal at state $(A^n p, p)$) for $n \in \mathbb{N}$ and $v^n + J^* = b^n = A^n p + Ap$.

4) $a^\infty = 1 + p$, $b^\infty = 1 + Ap$. Thus, $b^\infty > a^\infty$. Moreover, $b^\infty - c^\infty = p \cdot Ap \geq 0$. Thus, the action $(0, 1)$ is optimal at state $(p, 1)$ (and by symmetry, the action $(1, 0)$ is optimal at state $(1, p)$), and $v^\infty + J^* = b^\infty = 1 + Ap$.

Case 2: $p \in (\alpha_1, \tau]$

We show that the values of Lemma 15 satisfy the dynamic program of (41) and (42), by considering the four cases separately.

1) $a^* = b^* = 2$, $c^* = 2 - Ap$. Hence, either action $(0, 1)$ or $(1, 0)$ is optimal at state $(1, 1)$ and $v^* + J^* = a^* = b^* = 2$.

2) $a^0 = b^0 = p + Ap$ and $c^0 = 1$. Thus, $c^0 - a^0 = (1 - p)^2 - p \geq 0$ for $p \in (\alpha_1, \tau]$. Recall that τ is the root of $x = (1 - x)^2$. Hence, action $(1, 1)$ is optimal at state (p, p) and $v^0 + J^* = c^0 = 1$.

3) Consider $n \in \mathbb{N}$. $a^n = p + A^{n+1}p$ and $b^n = A^n p + Ap$. Thus, $b^n - a^n = p(1-p) \cdot A^{n-1}p \geq 0$. Moreover, $c^n = (1 - p)A^n p + p - Ap + 1$ and, thus, $b^n - c^n = 2Ap + p \cdot A^n p - p - 1 > 2Ap + p \cdot Ap - p - 1 = \varphi_1(p)$, which is positive for $p \in (\alpha_1, t]$. Thus, the action $(0, 1)$ is optimal at state $(p, A^n p)$ (and by symmetry, the action $(1, 0)$ is optimal at state $(A^n p, p)$) for $n \in \mathbb{N}$ and $v^n + J^* = b^n = A^n p + Ap$.

4) $a^\infty = 1 + p$, $b^\infty = 1 + Ap$. Thus, $b^\infty > a^\infty$. Furthermore, $c^\infty = 2 - Ap$ and $b^\infty - c^\infty = 2Ap - 1 = -\varphi_0(1 - p)$, which is positive for $p > 1 - \alpha_0$. Since $\alpha_1 > 1 - \alpha_0$, we have that for $p \in (\alpha_1, \tau]$, the action $(0, 1)$ is optimal at state $(p, 1)$ (and by symmetry, the action $(1, 0)$ is optimal at state $(1, p)$) and $v^\infty + J^\infty = b^\infty = 1 + Ap$.

Case 3: $p \in (\alpha_{m+1}, \alpha_m]$, $m \in \mathbb{N}$

We show that the values of Lemma 15 satisfy the dynamic program of (41) and (42), by considering each case separately. Recall $\zeta(x) = 1 + x^2 + x^3$.

1) $a^* = b^* = 2$ and $c^* = 2 - J^*$. Hence, either action $(0, 1)$ or $(1, 0)$ is optimal at state $(1, 1)$ and $v^* + J^* = a^* = b^* = 2$.

2) $a^0 = b^0 = p + w^1$ and $c^0 = 2p - p^2 J^* + (1-p^2)v^0$. Hence, $a^0 - c^0 = -p^2\varphi_0(p)/\zeta(p)$ which is positive for $p < \alpha_0$. Thus, for $p \in (\alpha_{m+1}, \alpha_m]$, either action $(0,1)$ or $(1,0)$ is optimal at state (p,p) and $v^0 + J^* = c^0$

3) Consider $n < m$. Then, $c^n - a^n = -pA^n p\varphi_0(p)/\zeta(p)$, which is positive for $p \in [0, \alpha_0]$ and thus for $p \in (\alpha_{m+1}, \alpha_m]$. Moreover, $c^n - b^n = -p^2\varphi_n(p)/\zeta(p)$ which is positive for $p \in (0, \alpha_n]$ and hence for $p \in (\alpha_{m+1}, \alpha_m]$. (Recall that α_n forms an decreasing sequence.). Thus, the action $(1,1)$ is optimal at state $(p, A^n p)$ for $n < m$ (and by symmetry for state $(A^n p, p)$ for $n < m$) and $v^n + J^* = c^n$.

4) Consider $n = m$. Then, $a^m = p + A^{m+1}p$ and $c^m - a^m = -(1-p)^m\varphi_1(p) + (1 - 2p - p^3)$. The first term is positive for $p \in [0, \alpha_m]$ and since the second term is larger than $\varphi_1(p)$, the second term is also positive in that interval. Moreover $c^m - b^m = -p^2\varphi_n(p)/\zeta(p)$ which is positive for $p \in [0, \alpha_m]$. Thus, both terms are positive for $p \in (\alpha_{m+1}, \alpha_m]$. Hence, the action $(1,1)$ is optimal at the state $(p, A^m p)$ (and by symmetry for the state $(A^m p, p)$) and $v^m + J^* = c^m$.

5) Consider $n > m$. Then, $a^n = p + A^{n+1}p$ and $b^n = A^n p + J^*$. Thus, $b^n - a^n = -p\varphi_0(p)/\zeta(p) - p(1-p)^{n+1} > -p\varphi_0(p)/\varphi - p(1-p) = -p^2\varphi_1(p)/\varphi$ which is positive for $p \in [0, \alpha_1]$ and the second term is always positive. Moreover, $b^n - c^n = p^2\varphi_n(p)/\zeta(p)$ which is positive for $p \in (\alpha_n, 1]$ and hence for $p \in (\alpha_{m+1}, \alpha_m] \subset (\alpha_n, 1]$. (Recall that α_n forms an decreasing sequence.). Thus, the action $(0,1)$ is optimal at state $(p, A^n p)$, for $n > m$ (and by symmetry the action $(1,0)$ is optimal at state $(A^n p, p)$ for $n > m$) and $v^n + J^* = b^n$.

6) $a^\infty = p + J^*$ and $b^\infty = 1 + w^1$. Thus, $b^\infty - a^\infty = -p\varphi_0(p)/\zeta(p)$, which is positive for $p < \alpha_0$. Moreover, $c^\infty = 1 + p - pJ^* + (1-p)v^0$ and, thus, $b^\infty - c^\infty = p^2(1 + (1-p)^2)/\zeta(p)$ which is always positive. Thus, for $p \in (\alpha_{m+1}, \alpha_m]$, the action $(0,1)$ is optimal at state $(p,1)$ (and by symmetry the action $(1,0)$ is optimal at state $(1,p)$) and $v^\infty + J^\infty = b^\infty = 2J^*$.